Ranking Interactions for a Curation Task

Clematide, S ; Rinaldi, Fabio

Abstract: One of the key pieces of information which biomedical text mining systems are expected to extract from the literature are interactions among different types of biomedical entities (proteins, genes, diseases, drugs, etc.). Different types of entities might be considered, for example protein-protein interactions have been extensively studied as part of the Bio Creative competitive evaluations. However, more complex interactions such as those among genes, drugs, and diseases are increasingly of interest. Different databases have been used as reference for the evaluation of extraction and ranking techniques. The aim of this paper is to describe a machine-learning based reranking approach for candidate interactions extracted from the literature. The results are evaluated using data derived from the Pharm GKB database. The importance of a good ranking is particularly evident when the results are applied to support human curators.

DOI: https://doi.org/10.1109/ICMLA.2011.119

Ranking interactions for a curation task

Simon Clematide, Fabio Rinaldi
Institute of Computational Linguistics
University of Zurich, Switzerland
siclemat@cl.uzh.ch,fabio.rinaldi@uzh.ch

***Abstract*—One of the key pieces of information which biomedical text mining systems are expected to extract from the literature are interactions among different types of biomedical entities (proteins, genes, diseases, drugs, etc.).**

Different types of entities might be considered, for example protein-protein interactions have been extensively studied as part of the BioCreative competitive evaluations. However, more complex interactions such as those among genes, drugs, and diseases are increasingly of interest. Different databases have been used as reference for the evaluation of extraction and ranking techniques.

The aim of this paper is to describe a machine-learning based reranking approach for candidate interactions extracted from the literature. The results are evaluated using data derived from the PharmGKB database. The importance of a good ranking is particularly evident in the case the results are applied to support human curators.

***Index Terms*—Text Mining, Literature Curation, Machine Learning, Maximum Entropy**

I. Introduction

The wealth of published information in the biomedical domain is at the same time an opportunity and a challenge. Accessing this information, and making sense of it, is an increasingly difficult task, which requires considerable expertise. In order to help the biologists to quickly locate the essential information that they need, different organizations provide curated databases, which organize the available knowledge about a particular specific subject, for example UniProt/SwissProt [1] is one of the most authoritative resources concerning proteins, BioGrid [2] is the broadest database describing gene and protein interactions.

Most reference databases are created and maintained using a very costly and expensive manual curation procedure, which involves highly skilled professionals. It has been observed already a few years ago that such an approach is not sufficiently efficient in order to cope with the increasing quantity of published results [3]. In order to support this process, researchers are turning their attention to text mining methodologies, not with the aim of replacing manual curation, which is not possible in the foreseeable future, but rather with the aim of providing tools that can make the curation process more efficient. Clearly such tools will need to be tailored to the specific task or database where they are going to be deployed, however some major tendencies are already clear and will shape the future development of the field. Some of the fundamental tasks that text mining systems are required to deal with are entity recognition, concept identification and interaction detection.

The text mining community has been organizing a number of shared tasks aiming at providing an infrastructure for the comparative evaluation of different text mining technologies. One such task which is of particular relevance to the work described in this paper is the protein-protein interaction task which has taken place in the 2006 and 2009 editions of the BioCreative competitive evaluations. The organizers provide a collection of annotated documents as training data (typically derived from one of the curated databases) and a separate collection of unannotated documents as test data. Participants have a limited time frame to process the training data and deliver results to the organizers, who will then score those results against a previously withheld gold-standard, using a set of metrics suited to the task.

In this paper we focus on a different type of interactions, namely those among genes, drugs and diseases, and we use information derived from the PharmGKB database [4], [5] as our gold standard. One advantage of the PharmGKB database is that it provides its data in a format which is structurally similar to the format used by the protein interaction task of BioCreative, thus allowing us to use the same tools for scoring our results. We propose a simple methodology to achieve a high-quality ranked list of candidate interactions starting from know entities and their normalized identifiers. Once entities have been identified and disambiguated, candidate interactions can be generated with simple techniques, for example co-occurrence within the same text span. However such candidates would be too numerous to be useful, so proper ranking techniques are necessary in order to render those results accessible and really useful for a curation task. This paper describes in particular a machine learning approach towards reranking of candidate interactions using a maximum entropy method. We conclude with a brief overview of an integrated curation system where the results described in the paper will be applied.

II. Text Mining Approach

In order to perform a simple and replicable experiment we refrain from sophisticated entity recognition approaches and do not use any external database of names and identifiers, and instead use only the terms and entities provided by PharmGKB itself, which can be downloaded in a simple textual format. These resources include the terms used in the curated papers and unique identifiers for each corresponding concept, in particular 30351 terms (2986 IDs) for drugs, 28633 terms (3198 IDs) for diseases, 176366 terms (28633 IDs) for genes.

Relationships are represented as binary interactions between two typed identifiers, with supporting evidence provided in the form of PubMed article ID referring to publications which mention the specific interaction. Our gold standard for all the experiments described in this paper is the set of interactions provided by PharmGKB.

For a number of relationships involving genetic polymorphisms, an additional reference to the Single Nucleotide Polymorphism database at NCBI (dbSNP)[1] is provided in the form of a rsID (reference single-nucleotide polymorphism [SNP] ID). Interactions that are recognized as playing an important role in a known pathway are additionally annotated with a reference to the specific pathway (which is described in a separate file). There are a total of 22827 interactions available in the version of PharmGKB which we have used for the experiments described in this paper. Once the multiple evidence sources for each interaction are separated, we obtain a total of 36557 triples consisting of two entity IDs and one source IDs. These triples can be classified according to the type of the evidence, which can be either a PubMed identifier (26122), a pathway reference (5467) or a rsID (4968).

In our experiments we consider only the interactions which are supported by a PubMed identifier, discarding the pathway-based and rsID-based interactions. These 26122 binary interactions, which are based upon 5062 distinct articles, can be used as a "gold standard" in a text mining task analogous to the protein-protein interaction task defined in the BioCreative text mining competitions [6], [7]. Participants to this task are asked to identify (by automated text processing) protein-protein interactions in a set of PubMed publications specified by the organizers. The organizers initially provide "training data" in the form of articles with known interactions, and in a subsequent phase the participants have to identify such interactions in a set of unseen articles, and deliver them to the organizers in a simple format (the UniProt IDs of the two proteins and the PubMed ID of the article). The organizers then score automatically the results of the participants against a manually identified set of correct interactions.

A. Evaluation Measures

The format of the relationship file provided by PharmGKB lends itself to easy transformation into a format equivalent to that used for the protein-protein interaction task of BioCreative II.5 [7]. Given a text mining tool which can produce a ranked list of gene/drug/disease interactions, it becomes then possible to score these results against the PharmGKB-derived data using a scoring tool provided by the BioCreative organizers.

The BioCreative scorer returns an evaluation of the results according to the standard metrics used in information retrieval (Precision, Recall, F-score) as well as a relatively novel measure called "AUC iP/R" (area under the curve of the interpolated precision/recall graph).[2] The purpose of the AUC iP/R measure (henceforth "AUC", not to be confused with the more frequently used "AUC of the ROC curve" metric) is to provide an indication of the quality of the ranking of the results. The intuitive idea is that, given equivalent P/R/F figures, correct predictions which occur towards the top of the ranked list of results are more useful than results which are lower in the ranking. The implicit assumption is that a curator could use the ranking to decide where to stop looking at the results, therefore a better ranking provides a better user experience. A recently proposed alternative measure of the ranking of the results is the "Threshold Average Precision" (TAP-k) [9], which (in slightly simplified terms) averages precision for the results above a given error threshold. The TAP-k metric is easier to interpret and directly relevant for the end user, who in most cases would not be willing to inspect a long list of results containing many false positives.

B. Text Processing

For our experiments, we automatically download from PubMed the abstracts corresponding to the PubMed IDs mentioned by the PharmGKB relationship file. All experiments described in this paper are based on this collection of abstracts. It would of course be desirable to work on full papers rather than abstracts, however not all these publications are freely downloadable, and most importantly, they are not available in a common format. The lack of a common format hinders the usability of full-text publications, as it makes it more difficult to identify significant zones of the papers (e.g. results sections) or zones that require special processing (e.g. tables).

We apply the OntoGene relation mining system (OG-RM) in order to annotate the input documents, using only the terminology provided by PharmGKB First, in a preprocessing stage, the input text is transformed into a custom XML format, and sentences and tokens boundaries are identified. The OntoGene pipeline also includes a step of term annotation and disambiguation [10], [11]. In order to account for possible surface variants a normalization step is included in the annotation procedure. The pipeline also includes part-of-speech taggers [12], a lemmatizer [13] and a syntactic chunker [14]. The rich annotations generated by the OntoGene pipeline can then be used to generate candidate interactions using a number of different criteria. Each token in the OntoGene annotation framework is assigned a unique identifier. Extracted terms can be related back to their position in text thanks to the unique token identifiers.

C. Relation Extraction

There are different ways in which the entities identified in each abstract can be combined, for example by co-occurrence in the same sentence, or by using a set of syntactic filters as done in our previous work [15], [16]. However, the approach which delivers the maximal recall is to generate all pairwise undirected combinations of ALL entities identified in the abstract. This approach can deliver a recall of slightly more than 60%, which is quite good considering that only abstracts

[1] http://www.ncbi.nlm.nih.gov/projects/SNP/

[2] The AUC iP/R curve is defined in [8], a detailed operative description of AUC iP/R, as used in the BioCreative evaluations, can be found at http://www.biocreative.org/tasks/biocreative-ii5/biocreative-ii5-evaluation/

are used.[3] However, this approach will massively overgenerate, therefore ranking of the results becomes absolutely necessary.

An initial ranking of the candidate interactions can be generated only on the basis of frequency of occurrence of the respective entities/terms:

$score(c_1, c_2) = (f(c_1) + f(c_2))/f(C)$

where $f(c_1)$ and $f(c_2)$ are the number of times the identifiers c_1 and c_2 are observed in the abstract, while $f(C)$ is the total count of all identifiers in the abstract. Once a score is assigned to each candidate pair, it is possible to filter out the most unlikely candidates, either by setting a threshold value for the score, or by selecting only the N-best candidates. Using one of these methods will result into variable values of Precision, Recall and F-score, depending on the exact value of the score threshold, or N parameter.

We know from our own previous experiments [15] that giving a "boost" to the entities contained in the title can produce a measurable improvement of ranking of the results (measured by the AUC or TAP metrics). We have empirically verified that the best value of such a boost is around 10. This is equivalent to count ten times the entities in the title, or in other words to treat the title as if it were repeated ten times.

The approach described above will be referred to as **art** in the rest of this paper.

The ranking of relation candidates using a simple frequency based confidence score can be further optimized if we apply a supervised machine learning method. This approach will be referred to as **art-me** in the rest of this paper. First we automatically identify the noisy concepts that our term recognizer generates in order to penalize them. Second, we need to adapt to highly-ranked false positive relations which are generated by our frequency based approach. The goal is to identify some global preference oder biases which can be found in the PharmGKB relational database. One technique is to weight individual concepts according to their likeliness to appear as an entity in a gold relation. Another technique is to generally penalize the score of relations between concepts of the same type.

For a precise description of our optimized ranking approach, we need to introduce some notation. In the following, the notation t refers to a standardized form of a term as it was recognized by our term recognizer. The standardization step currently consists of down-casing and removal of some punctuation characters (hyphens and parentheses) and is mainly motivated by the need to reduce data sparseness problems: for instance, *"Fc (gamma) - receptor"* is standardized to *"fc gamma receptor"*. Our term recognizer aggressively modifies term names (i.e. removes material from an entry in the ontology or creates on-the-fly acronyms) while matching terms. For instance, the term form 'neuronal' may be identified as gene concepts `PA134898200`, `PA134924203`, `PA134896732` because they have "neuronal protein" as one of their designator. The combination of a term t and one of its valid concept groundings c is noted as $t : c$. When we count the occurrences of a term-concept combination we apply a cap of 6 (in order to reduce data sparseness) to the raw article frequency $f(t : c)$:

$$C(t:c) = \begin{cases} 6 & \text{if } f(t:c) \geq 6 \\ f(t:c) & \text{otherwise.} \end{cases}$$

Next we define a predicate $gold(c)$ which is true for an article A if there exists at least one relation for A in the PharmGKB gold standard where concept c appears. Using the notions defined beforehand, we can specify the probability of concept c taking part in at least one gold relation given the concept c, a term form t, and its combination count $C(t : c)$:

$$P(\, gold(c) = 1 \mid c, t, C(t:c)\,)$$

The relevance score of a concept c for an article A is given by:

$$score(c) = \sum_{t:c \in A} C_z(t:c) \times P(\, gold(c) = true \mid c, t, C(t:c)\,)$$

The expression $C_z(t : c)$ designates the zoned occurrence counts of $t : c$ with a boosting factor for occurrences in the title zone. For the abstracts at hand, a boost factor of 10 was empirically verified as a good setting.

Having determined the individual score of each concept c, we can basically add up this information to score the relation between two concepts in an article:

$$relscore(c_1, c_2) = (score(c_1) + score(c_2)) \times penalty(c_1, c_2)$$

The $penalty$ currently affects only relations between concepts of the same type, which are underrepresented in the PharmGKB:

$$penalty(c_1, c_2) = \begin{cases} 0.1 & \text{if both concepts have the same type} \\ 1 & \text{otherwise.} \end{cases}$$

We estimate $P(\, gold(c) = 1 \mid c, t, C(t:c)\,)$ with the help of a Maximum Entropy (ME) optimization tool (megam [17]) using the output from our term recognizer applied to a training set of PharmGKB abstracts. The training set is a randomly selected 90% subset of the full set of abstract mentioned in the PharmGKB database. We use $c, t, C(t : c)$ as a joint feature for the maximum entropy classifier and the value of $gold(c)$ as its binomial class. In order to reduce sparse data problems we introduce a smoothing method by also adding all features with lower frequencies than $C(t : c)$.

The Maximum Entropy Classifier computes the probability of each concept by using the weights in its exponential model, which also includes an apriori bias according to the global distribution of class 0 and 1, where class 1 express correct (gold standard) relations.

The maximum entropy classifier assigns to unseen terms (t - terms not present in the training data) a default probability based on the distribution of the training instances. However,

[3] These values represent the recall using only the textual information in the title and abstract. For the results presented further on we also added some of the metadata (MeSH terms and chemical substances) which leads to a maximum recall of 69% on the evaluation data set.

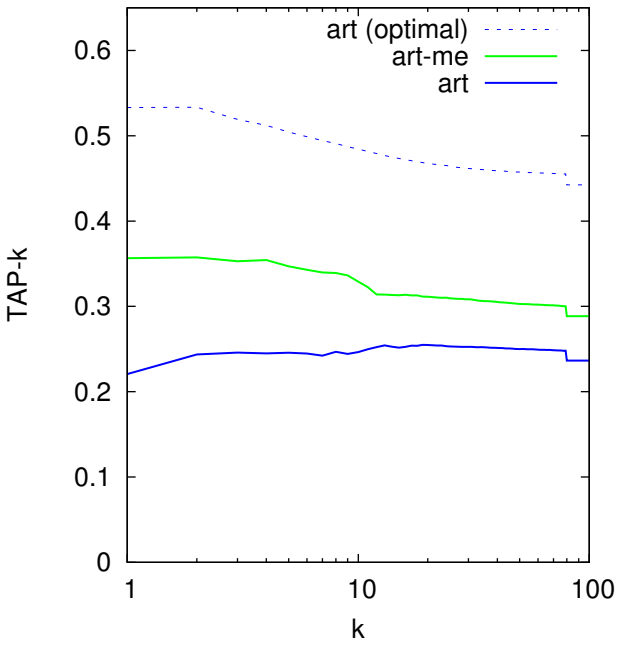

Fig. 2. The TAP-k values for our different approaches on the 10% evaluation data set. The horizontal axis shows the k threshold. The vertical axis shows the resulting TAP for a given k. Note that the flat segment is due to the padding of the result list with dummy results if too few results are reported for a query to reach k. The dotted lines show the TAP-k values which could be achieved if all true positive hits of the system would be optimally ranked as hits with the highest confidence score.

we can specify better backoff probabilities if we take into account the concept(s) c of term t. Our current backoff model works as follows: if the concept c of an unseen t is known, the average of all known term-concept combination probabilities is used. All concepts of the evaluation set were known from the training set, therefore this backoff model was sufficient for the data at hand.

Meth.	Docs	TP	FP	FN	AUCiP/R	n
art	478	194	284	1570	0.246	1
art	478	292	660	1472	0.301	2
art	478	349	1076	1415	0.327	3
art	478	428	1923	1336	0.348	5
art	478	542	4061	1222	0.371	10
art	478	884	63104	880	0.391	all
art-me	478	283	195	1481	0.345	1
art-me	478	401	551	1363	0.418	2
art-me	478	466	959	1298	0.444	3
art-me	478	561	1790	1203	0.471	5
art-me	478	672	3931	1092	0.491	10
art-me	478	884	63104	880	0.507	all

TABLE I

RESULTS ON THE 10% EVALUATION DATA SET, CONTAINING A TOTAL OF 485 DOCUMENTS. THE FIRST COLUMN GIVES THE APPROACH USED. THE SECOND COLUMN REPORTS THE NUMBER OF DOCUMENTS WITH A LEAST ONE RESPONSE HIT. THE THIRD TO THE FIFTH COLUMN GIVE TRUE POSITIVES (TP), FALSE POSITIVES (FP) AND FALSE NEGATIVES (FN). THE SIXTH COLUMN CONTAINS THE MACRO AVERAGED AUCIP/R. THE SEVENTH COLUMN CONTAINS THE CUT-OFF VALUE n USED BY THE BIOCREATIVE EVALUATION TOOL AS A THRESHOLD ON THE NUMBER OF RESPONSE HITS WHEN COMPUTING THESE RESULTS. IN ROWS WITH $n = all$ NO THRESHOLD WAS APPLIED.

For a systematic evaluation using the supervised methods describe before, we split the corpus into 90% training data (4540 articles) and 10% test data (505 articles). Because the relation types are distributed unevenly over all documents, we tried to ensure an approximately similar distribution of different relation types in the two data sets.

Table I compares the performance of our two approaches as computed by the BioCreative evaluation tool with increasing cut-off thresholds n. As n is increased, more noise is allowed to appear. Note that this tool ignores gold standard annotations for documents where no response hits are generated by the evaluated system. The Maximum Entropy ranking (art-me) achieves a substantial improvement in terms of AUC iP/R. Figure 1 visualizes the same findings as performance curves in terms of precision, recall and F-Score. The high impact of recall on AUCiP/R is obvious in these plots. In Figure 2 we report the performance of the same approaches as above but using the TAP-k metric. The adapted ranking using Maximum Entropy optimization leads to a very similar degradation curve of the TAP score with increasing k threshold, compared to the idealized curve computed by ranking all true positives higher than the false positives. This indicates that our optimized ranking performs well not only on the top scored results. However, most impressing is that we can increase TAP-1 from 0.22 to 0.40 which is an improvement of 182% (compare 'art' with 'art-me' in Figure 2, solid lines). The high value of TAP-1 indicates that correct relations really do tend to appear on top of the ranking.

As a continuation of this work, we would like to explore the possibility of providing an indicative prediction of the number of interactions to be found in a paper on the basis of the textual content of the paper, possibly taking into account the initial steps of interaction with a curator. Being able to provide such indication before or at the initial stages of the curation process would help the curators to decide at which point in the curation process it is most sensible to stop after having found a given number of correct interactions. This is particularly relevant because documents differ greatly in the number of interactions they describe, ranging from just one to several hundreds in a few documents describing high-throughput experiments. In PharmGKB we have observed that 40% of the documents contain only one relation, however they contribute less than 10% of all relations. Approx. 90% of the documents contain 10 or less relations, however these documents contain less than 50% percents of all relations. So the remaining 10% of documents (which contributes more than 50% to the relations) has a much higher number of relations per document.

III. USAGE IN A CURATION ENVIRONMENT

Advanced text mining techniques are now reaching a maturity level that renders them increasingly relevant for the process of curation of biomedical literature. As part of our own research in this area we developed a curation system called "OntoGene Document INspector" (ODIN [18]) which interfaces with our text mining pipeline. We have used a version of ODIN for our participation to the 'interactive

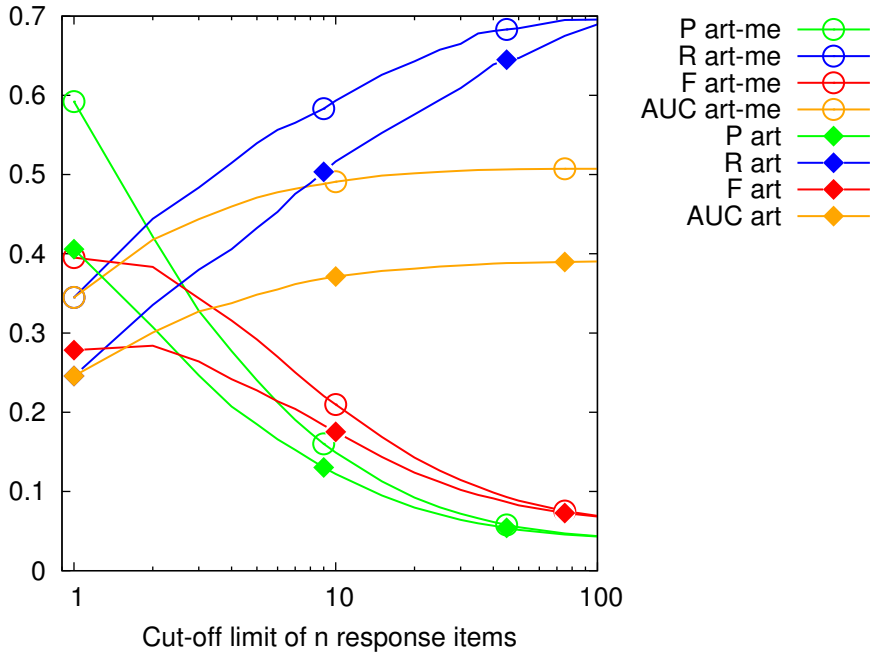

Fig. 1. The results from the BioCreative evaluation tool for our different approaches on the 10% evaluation data set. The horizontal axis shows the cut-off value limiting the number of hits that are evaluated by the tool. The vertical axis shows macro averaged results of precision (P), recall (R), F-score (F) and AUC iP/R for our different approaches. Note that these results were computed by ignoring documents without hits in the system responses (this is the default setting for the BioCreative evaluations). See Table I for the number of documents that produce hits.

curation' task (IAT) of the BioCreative III competition [19]. This was an informal task without a quantitative evaluation of the participating systems. However, the curators who used the system commented extremely positively on its usability for a practical curation task.

More recently, we have created a version of ODIN which allows inspection of abstracts automatically annotated with PharmGKB entities (the annotation is performed using the Ontogene pipeline). Users can access either preprocessed documents, or enter any PubMed identifier and have the corresponding abstract processed "on the fly". For the documents already in PharmGKB it is also possible to inspect the gold standard and compare the results of the system against the gold standard. The curator can inspect all entities annotated by the system, and easily modify them if needed (removing false positives with a simple click, or adding missed terms if necessary). The modified documents can be sent back for reprocessing if desired, obtaining therefore modified candidate interactions. The user can also inspect the set of candidate interactions generated by the system, and act upon them just as on entities, i.e. confirm those which are correct, remove those which are incorrect. Candidate interactions are presented ordered according to the score which has been assigned to them by the text mining system, therefore the curator can choose to work with only on a small set of highly ranked candidates, ignoring all the rest (see Figure 3). Recent user experiments using our curation environment which makes use of the ranking proposed by the method described above have shown encouraging results.

IV. Conclusions

In this paper we have presented a maximum entropy approach towards the reranking of candidate interactions obtained from a simple text mining approach, which can considerably enhance the usability of a curation environment. We have shown how it is possible to use existing tools to score the results and provide reliable metrics, including not only the traditional Precision, Recall and F-score but also the increasingly important measures of ranking quality, such as "AUC iP/R" or "TAP-k".

We have presented our own approach towards the mining of pharmacogenomics relationships and scored it against the PharmGKB dataset. Our experiments show that this task is feasible, and our results might offer a useful baseline for further developments in this area. Finally, we have presented an implementation of our assisted curation environment (ODIN) specifically adapted to the PharmGKB dataset.

V. Acknowledgements

This research is partially funded by the Swiss National Science Foundation (grant 100014-118396/1). Additional support is provided by Novartis Pharma AG, NITAS, Text Mining Services, CH-4002, Basel, Switzerland.

References

[1] UniProt Consortium, The universal protein resource (uniprot), Nucleic Acids Research 35 (2007) D193–7.

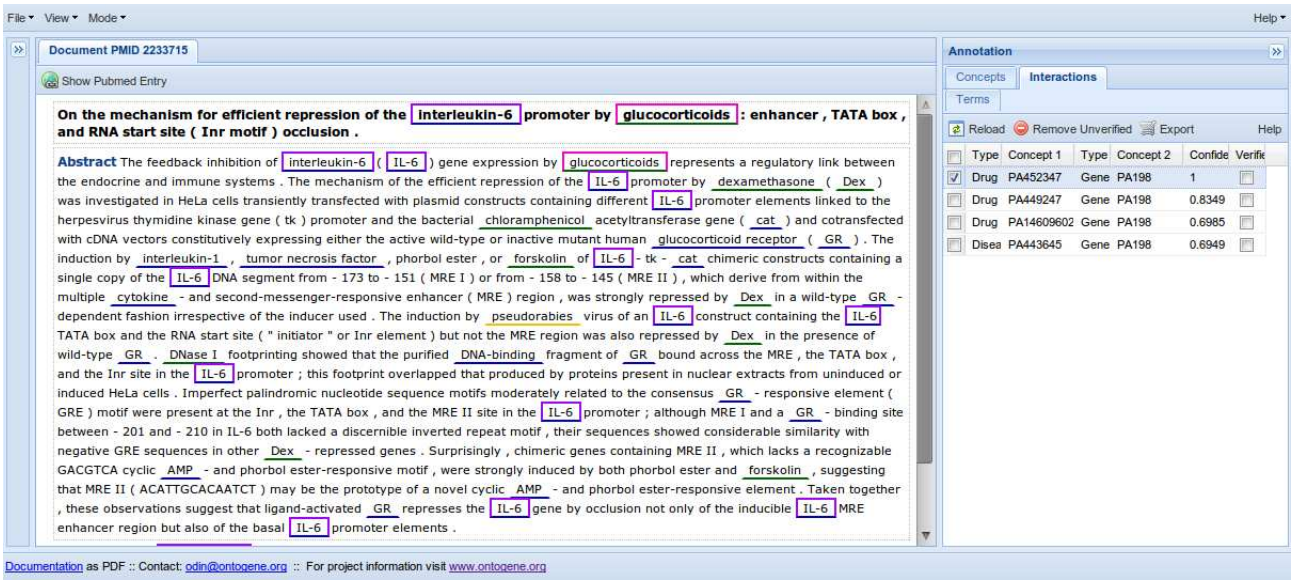

Fig. 3. Example of interaction with the ODIN system. Terms identified by the system are underlined in the abstract. Candidate interactions are shown in the left-hand-side panel. Selecting an interaction automatically highlights the terms in the document which correspond to the entities in the interaction.

[2] C. Stark, B.-J. Breitkreutz, T. Reguly, L. Boucher, A. Breitkreutz, M. Tyers, Biogrid: A general repository for interaction datasets, Nucleic Acids Research 34 (2006) D535–9.

[3] W. A. Baumgartner, K. B. Cohen, L. M. Fox, G. Acquaah-Mensah, L. Hunter, Manual curation is not sufficient for annotation of genomic databases, Bioinformatics 23 (13) (2007) i41–48. doi:10.1093/bioinformatics/btm229.
URL http://bioinformatics.oxfordjournals.org/cgi/content/abstract/23/13/i41

[4] T. Klein, J. Chang, M. Cho, K. Easton, R. Fergerson, M. Hewett, Z. Lin, Y. Liu, S. Liu, D. Oliver, D. Rubin, F. Shafa, J. Stuart, R. Altman, Integrating genotype and phenotype information: An overview of the pharmgkb project, The Pharmacogenomics Journal 1 (2001) 167–170.

[5] K. Sangkuhl, D. S. Berlin, R. B. Altman, T. E. Klein, Pharmgkb: Understanding the effects of individual genetic variants, Drug Metabolism Reviews 40 (4) (2008) 539–551, pMID: 18949600. doi:10.1080/03602530802413338.
URL http://informahealthcare.com/doi/abs/10.1080/03602530802413338

[6] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, A. Valencia, Overview of the protein-protein interaction annotation extraction task of BioCreative II, Genome Biology 9 (Suppl 2) (2008) S4.

[7] F. Leitner, S. A. Mardis, M. Krallinger, G. Cesareni, L. A. Hirschman, A. Valencia, An overview of biocreative ii.5, IEEE/ACM Transactions on Computational Biology and Bioinformatics 7 (3) (2010) 385–399. doi:10.1109/TCBB.2010.50.

[8] C. D. Manning, P. Raghavan, H. Schütze, Introduction to Information Retrieval, Cambridge University Press, 2008.

[9] H. D. Carroll, M. G. Kann, S. L. Sheetlin, J. L. Spouge, Threshold average precision (tap-k): a measure of retrieval designed for bioinformatics, Bioinformatics 26 (14) (2010) 1708–1713. doi:10.1093/bioinformatics/btq270.

[10] F. Rinaldi, K. Kaljurand, R. Saetre, Terminological resources for text mining over biomedical scientific literature, Journal of Artificial Intelligence in Medicine 52 (2) (2011) 107–114.

[11] K. Kaljurand, F. Rinaldi, T. Kappeler, G. Schneider, Using existing biomedical resources to detect and ground terms in biomedical literature, in: Proceedings of the 12th Conference on Artificial Intelligence in Medicine (AIME09), 2009, pp. 225–234.

[12] Y. Tsuruoka, Y. Tateishi, J.-D. Kim, T. Ohta, J. McNaught, S. Ananiadou, J. Tsujii, Developing a robust part-of-speech tagger for biomedical text, in: Advances in Informatics - 10th Panhellenic Conference on Informatics, LNCS 3746, 2005, pp. 382–392.

[13] G. Minnen, J. Carroll, D. Pearce, Applied morphological processing of English, Natural Language Engineering 7 (3) (2001) 207–223.

[14] A. Mikheev, S. Finch, A workbench for finding structure in texts, in: Proceedings of the Fifth Conference on Applied Natural Language Processing, Association for Computational Linguistics, Washington, DC, USA, 1997, pp. 372–379. doi:10.3115/974557.974611.
URL http://www.aclweb.org/anthology/A97-1054

[15] F. Rinaldi, T. Kappeler, K. Kaljurand, G. Schneider, M. Klenner, S. Clematide, M. Hess, J.-M. von Allmen, P. Parisot, M. Romacker, T. Vachon, OntoGene in BioCreative II, Genome Biology 9 (Suppl 2) (2008) S13. doi:10.1186/gb-2008-9-s2-s13.
URL http://genomebiology.com/2008/9/S2/S13

[16] F. Rinaldi, G. Schneider, K. Kaljurand, S. Clematide, T. Vachon, M. Romacker, OntoGene in BioCreative II.5, IEEE/ACM Transactions on Computational Biology and Bioinformatics 7 (3) (2010) 472–480. doi:10.1109/TCBB.2010.50.

[17] H. Daumé III, Notes on CG and LM-BFGS optimization of logistic regression, paper available at http://pub.hal3.name#daume04cg-bfgs, implementation available at http://hal3.name/megam/ (August 2004).

[18] F. Rinaldi, S. Clematide, G. Schneider, M. Romacker, T. Vachon, ODIN: An advanced interface for the curation of biomedical literature, in: Biocuration 2010, the Conference of the International Society for Biocuration and the 4th International Biocuration Conference., 2010.

[19] C. Arighi, P. Roberts, S. Agarwal, S. Bhattacharya, G. Cesareni, , r. a. Chatr, S. Clematide, P. Gaudet, M. G. Giglio, I. Harrow, E. Huala, M. Krallinger, U. Leser, D. Li, F. Liu, Z. Lu, L. Maltais, N. Okazaki, L. Perfetto, F. Rinaldi, R. Saetre, D. Salgado, P. Srinivasan, P. E. Thomas, L. Toldo, L. Hirschman, C. H. Wu, Biocreative iii interactive task: an overview, BMC Bioinformatics, special issue on BioCreative III - (2011) –, (accepted for publication).